AISRP Annual Report: September 28, 2005

Segmented Nonparametric Models of Distributed Data: From Photons to Galaxies

Jeffrey D. Scargle

Jeffrey.D.Scargle@nasa.gov Space Science and Astrobiology Division, NASA Ames Research Center

Michael Way, Co-Investigator

Pasquale Temi, Co-Investigator

Abstract: A novel technique, and an efficient algorithm to implement it, provides piecewise-constant models for a variety of one dimensional data types common in high energy astrophysics and cosmology, as well as in physical simulations of astrophysical systems. We have improved this algorithm in various ways (speed, parameter selection, alternative fitness functions, etc.) However, the major accomplishment was the extension of this methodology to 2D data (such as galaxy surveys), 3D data (redshift surveys), and higher dimensional data spaces. We discovered a way to transform higher dimensional problems into approximately equivalent 1D cases, solvable with the 1D algorithm. A major mathematical goal of this work, to find rigorous solutions in higher dimensions, has not yet been accomplished. We have applied the methods to GLAST photon maps, 3D galaxy positions derived from redshift surveys, and stellar infrared data useful for detection of molecular clouds. Ultimately these methods will be applicable to a wide range of astrophysical problems.

1. The Problem

- A huge volume of data.
- The desire to make scientific sense out of same.

. . . a summary of one of the most difficult problems universally facing astrophysicists. It is a cliché that the size and complexity of data from modern astrophysical programs is growing exponentially. Cliché or not, it implies the need for new tools covering the span from organized data acquisition and archiving to automatic–or at least semiautomatic–reduction and analysis software.

In this research we have focused on the problem of estimating a function $Y(x)$–that is, a physical quantity Y as a function of an independent variable x. Examples: the radiant flux of an astronomical object as a function of time; the density of galaxies as a function of position in the Universe; or the intensity of the cosmic microwave background as a function of position on the sky.

2. The Solution

We have a modest proposal for a solution to at least part of the above problem. We have developed an objective procedure for starting with a possibly multivariate set of data, distributed over a data space of almost any sort, and deriving an easily interpretable representation of the underlying physical process.

2.1. The Data

The data input to our procedure must be in the form of measurements related to the dependent variable Y for a set of values of the

independent variable x. We say "related to" because the connection between the measurements and Y may be indirect. For example, the measurement of the position of a galaxy on the sky, and of its redshift, are not a direct measurement of the density of galaxies, but indirectly lead to an estimate of this quantity.

We have been vague about the domain of possible values of x, since this aspect of the data also exhibits a great deal of variety. One may have measurements of a quantity at a point, or indeed the position of a point. One may have a measurement over a set of small regions distributed regularly or randomly, or some other way, over a region of a smooth space. One may have measurements over small regions that form a partition of the data space, or even sparsely over the space.

From these comments it seems that the observational setup can be almost arbitrary. But we will see in the next section that the nature of the representation, or *model* for Y, and the way our scheme estimates it, place some restrictions on the data structure.

2.2. The Model

The model is simple: Y is assumed to be constant over a set of regions of the data space. These regions and the single number giving the corresponding value of Y make up a structure we call a *block*. In other words, we adopt a piecewise constant model. The aspects of the model that need to be determined from the data are specifications of the constant value on each block and of the locations of the blocks.

3. The 1D Problem

The one dimensional version of the measurement process and its modeling has been discussed in several earlier papers, including one that presents an efficient algorithm that exactly solves the estimation problem. This paper is posted on the AISRP PI webpage

`aaaprod.gsfc.nasa.gov/aisrp/public/ProjectList.cfm`

The idea of applying dynamic programming to this problem is not novel, and seems to have been first carried out by Richard Bellman in the 1950s. The novelty of our paper is that our version of the algorithm automatically finds the number of elements in the optimal partition, whereas all previous published algorithms took this value as given.

4. The 2D and Higher Problem

We have shown how a simple approach allows the optimal partition in data spaces of higher dimensions to be obtained, using the 1D algorithm described in the previous section. The details are in a paper posted in draft form at `astrophysics.arc.nasa.gov/~jeffrey/`

5. Applications

The basic philosophy behind *Bayesian Blocks*, and the algorithm implementing it in 1D problems, has made its way into a variety of applications. Here is a summary of those known as of this date. In all of the URLs below, *http://* is understood.

Astrophysics Source Code Library: code, documentation (also available through the Physical Sciences Information Gateway)

- `ascl.net/block.html`

Chandra X-ray Observatory data analysis (notable Michael Nowak and Fred Baganoff)

- space.mit.edu/CXC/analysis/SITAR/bb_experiment.html
- space.mit.edu/CXC/analysis/SITAR/functs_bb.html
- online.itp.ucsb.edu/online/galactic_c05/baganoff/pdf/baganoff.pdf
- www.edpsciences.org/articles/aa/pdf/2004/43/aa0495-04.pdf
- www.astrostatistics.psu.edu/datasets/Chandra_flares.html
- hea-www.cfa.harvard.edu/CHAMP/RESULTS_PAPERS/paperI_full.pdf
- xxx.lanl.gov/pdf/astro-ph/0310567
- www.journals.uchicago.edu/cgi-bin/resolve?id=doi:10.1086/379819

Other X-ray Observatories: USA (Unconventional Stellar Aspect X-ray telescope), XMM-Newton

- www.aas.org/publications/baas/v32n3/head2000/171.htm
- aanda.u-strasbg.fr:2002/papers/aa/full/2002/35/aah3063/node2.html

GLAST

- glast.gsfc.nasa.gov/science/grbst/workplans/A6_reqs_DLB.pdf
- www-glast.slac.stanford.edu/software/Workshops/January01Workshop/talks

Swift

- www.swift.ac.uk/BAT_GSW_Manual_v2.pdf

Stellar Variability

- www.mpa-garching.mpg.de/~cosmo/hambaryan.ps.gz
- www.aip.de/People/ASchwope/papers/aah3063.pdf
- www.astro.psu.edu/coup/COUP_Suns.pdf

Compton GRO/BATSE

- wwwgro.unh.edu/bursts/cgrbnewb.html
- www.batse.msfc.nasa.gov/events/5hgrbs/program/abstracts/G-07/G-07.html Istvan Horvath
- arxiv.org/pdf/astro-ph/0507016

Solar flare applications; Michael Wheatland, Y.-J. Moon, and the Harvard-CfA group

- www.publish.csiro.au/paper/AS04062.htm
- www.lpi.usra.edu/meetings/ppv2005/pdf/8430.pdf
- solar.njit.edu/preprints/moon2002c.pdf
- www.journals.uchicago.edu/cgi-bin/resolve?id=doi:10.1086/421261

Machine Learning

- portal.acm.org/citation.cfm?id=956808
- www.ics.uci.edu/~pazzani/Publications/ssdb99.ps
- www.cc.gatech.edu/~isbell/reading/papers/p493-chiu.pdf

Related Mathematical Development

- wwwgro.unh.edu/bursts/cgrbnewb.html
- www.newton.cam.ac.uk/preprints/NI99014.pdf
- fasolt.openlib.org/~arclis/cache_konz/rclis/dbl/comrer/(2003)%253C%253
  www.ucy.ac.cy/ fanis/Papers/biometrika2001.pdf+
- www.cs.ucr.edu/~eamonn/pakdd200_keogh.ps

Gravitational Wave Astronomy

- www.astro.gla.ac.uk/users/matthew/firstYearReport.ps

Classes

- www.astro.cornell.edu/~cordes/A523/

Note: in the wavelet literature, there is something called Bayesian Block Shrinkage – similar, but the changepoints are not adaptive (data driven), instead are confined to lie on the wavelet scale hierarchy. Hence this literature is not included under **Related Mathematical Development**.